### 分析流程

### 分析步骤 1. 通过训练集数据来建立随机森林分类模型。 2. 通过建立的随机森林来计算特征重要性。 3. 将建立的随机森林分类模型应用到训练、测试数据，得到模型的分类评估结果。 4. 由于随机森林中具有随机性，每次运算的结果不一样，若保存本次训练模型，后续可以直接上传数据代入到本次训练模型进行计算分类。 5. 注：随机森林无法像传统模型一样得到确定的方程，通常通过测试数据分类效果来对模型进行评价。

### 详细结论

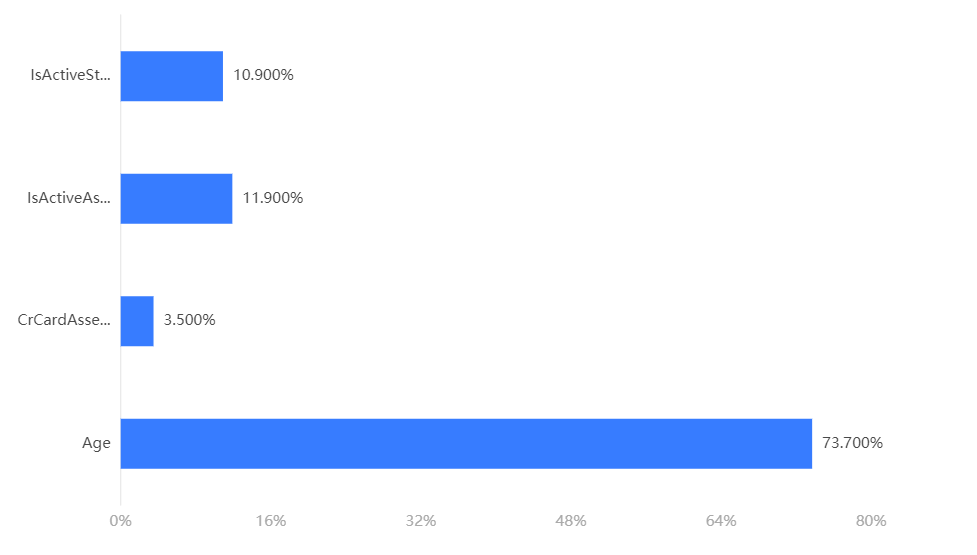
**输出结果1：模型参数**

|  |  |
| --- | --- |
| 参数名 | 参数值 |
| 训练用时 | 0.304s |
| 数据切分 | 0.9 |
| 数据洗牌 | 否 |
| 交叉验证 | 否 |
| 节点分裂评价准则 | gini |
| 决策树数量 | 100 |
| 有放回采样 | true |
| 袋外数据测试 | false |
| 划分时考虑的最大特征比例 | auto |
| 内部节点分裂的最小样本数 | 2 |
| 叶子节点的最小样本数 | 1 |
| 叶子节点中样本的最小权重 | 0 |
| 树的最大深度 | 10 |
| 叶子节点的最大数量 | 50 |
| 节点划分不纯度的阀值 | 0 |

**图表说明：**

上表展示了模型各项参数配置以及模型训练时长。

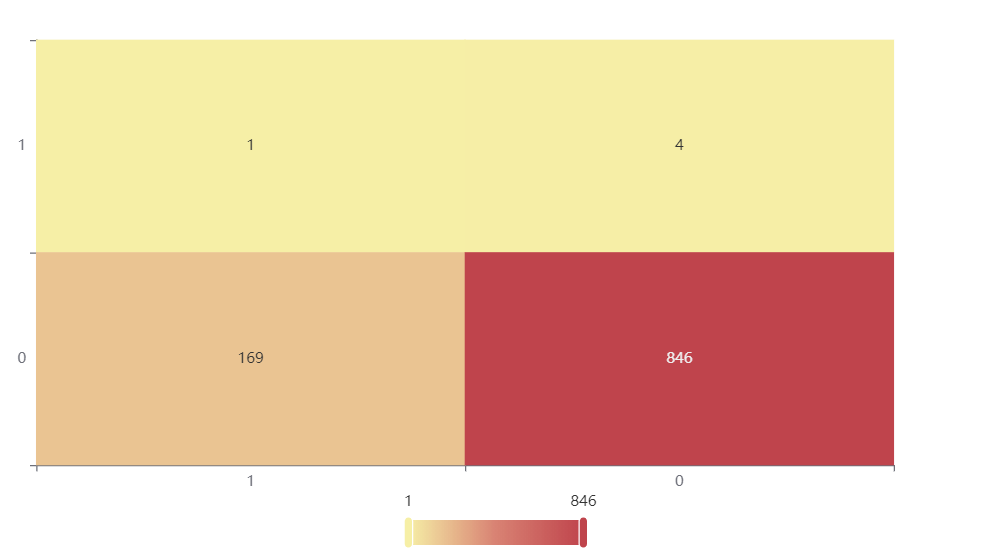
**输出结果2：特征重要性**



**图表说明：**

上柱形图或表格展示了各特征（自变量）的重要性比例。

**输出结果3：混淆矩阵热力图**



**图表说明：**

上表以热力图的形式展示了混淆矩阵。

**输出结果4：模型评估结果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 准确率 | 召回率 | 精确率 | F1 |
| 训练集 | 0.839 | 0.839 | 0.825 | 0.81 |
| 测试集 | 0.83 | 0.83 | 0.99 | 0.903 |

**图表说明：**

上表中展示了训练集和测试集的分类评价指标，通过量化指标来衡量随机森林对训练、测试数据的分类效果。  
● 准确率：预测正确样本占总样本的比例，准确率越大越好。  
● 召回率：实际为正样本的结果中，预测为正样本的比例，召回率越大越好。  
● 精确率：预测出来为正样本的结果中，实际为正样本的比例，精确率越大越好。  
● F1：精确率和召回率的调和平均，精确率和召回率是互相影响的，虽然两者都高是一种期望的理想情况，然而实际中常常是精确率高、召回率就低，或者召回率低、但精确率高。若需要兼顾两者，那么就可以用F1指标。  
● oob\_score：对于分类问题，oob\_score是袋外数据的准确率。若在建立树过程中选择有放回抽样时，大约1/3的记录没有被抽取。没有被抽取的自然形成一个对照数据集，可用于模型的验证。所以随机森林不需要另外预留部分数据做交叉验证，其本身的算法类似交叉验证，而且袋外误差是对预测误差的无偏估计（当算法参数选择了“袋外测试数据”后，才会通过oob\_score来检验模型的泛化能力）。

**输出结果5：测试数据预测评估结果**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 预测结果Y | Exited | 预测结果概率\_0.0 | 预测结果概率\_1.0 | IsActiveStatus | IsActiveAssetStage | CrCardAssetStage | Age |
| 0 | 0 | 0.9106697025634223 | 0.0893302974365775 | 1 | 0 | 0 | 34 |
| 0 | 0 | 0.7328094813950821 | 0.2671905186049182 | 3 | 6 | 6 | 56 |
| 0 | 1 | 0.6243078246600751 | 0.37569217533992494 | 1 | 2 | 9 | 43 |
| 0 | 0 | 0.7850127338142185 | 0.21498726618578134 | 0 | 2 | 9 | 38 |
| 1 | 1 | 0.4029040475830763 | 0.5970959524169238 | 2 | 1 | 7 | 50 |
| 0 | 0 | 0.5590212174768847 | 0.44097878252311523 | 5 | 8 | 2 | 47 |
| 0 | 0 | 0.8321741558755249 | 0.16782584412447504 | 0 | 2 | 9 | 34 |
| 1 | 0 | 0.44325271844804326 | 0.5567472815519569 | 2 | 2 | 9 | 46 |
| 0 | 1 | 0.7430377360199014 | 0.2569622639800987 | 0 | 2 | 2 | 40 |
| 0 | 0 | 0.7149957939256096 | 0.28500420607439053 | 2 | 1 | 2 | 42 |
| 0 | 0 | 0.8699273534296157 | 0.1300726465703847 | 3 | 8 | 9 | 37 |
| 0 | 0 | 0.7808444433292575 | 0.21915555667074205 | 4 | 8 | 9 | 42 |
| 0 | 0 | 0.8967023885177757 | 0.10329761148222437 | 5 | 8 | 9 | 74 |
| 0 | 0 | 0.8910305935283059 | 0.10896940647169417 | 4 | 8 | 9 | 36 |
| 0 | 1 | 0.8787306609708248 | 0.12126933902917511 | 3 | 7 | 7 | 63 |

**图表说明：**

上表格为预览结果，只显示部分数据，全部数据请点击下载按钮导出。  
上表展示了随机森林模型对测试数据的分类结果，分类结果值是拥有最大预测概率的分类组别。

**输出结果6：模型预测与应用**

**图表说明：**

● 系统会自动保存模型，需要注意的是：在机器学习中的随机森林算法保存的模型是非常复杂的，不是类似于线性回归那样可以用一个公式保存，系统以二进制文件方式进行序列化保存。  
● 由于随机森林具有随机性，每次训练的模型可能不一致，若保存本次训练模型，后续可以直接上传数据代入到本次训练模型进行计算预测。  
● 若删除本分析报告将会直接删除模型的缓存。

### 参考文献 [1] Scientific Platform Serving for Statistics Professional 2021. SPSSPRO. (Version 1.0.11)[Online Application Software]. Retrieved from https://www.spsspro.com. [2] 周志华. 机器学习[M]. 清华大学出版社, 2016.